DOCUMENT RESUME

ED 385 580 TM 024 018

AUTHOR Donoghue, John R.
TITLE An Empirical Examination of the IRT Information in Polytomously Scored Reading Items.
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY National Center for Education Statistics (ED), Washington, DC.
REPORT NO ETS-RR-93-12
PUB DATE Mar 93
NOTE 33p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Item Response Theory; *Multiple Choice Tests; *Reading Tests
IDENTIFIERS Empirical Research; *Information Function (Tests); National Assessment of Educational Progress; Partial Credit Model; Polytomous Items; *Polytomous Scoring; Short Answer Tests; Weighting (Statistical)

ABSTRACT

One question about polytomous items (which yield responses that can be scored as ordered categories) concerns how much information such items yield? Using the generalized partial credit item response theory (IRT) model, polytomous items from the 1991 field test of the National Assessment of Educational Progress Reading Assessment were calibrated with multiple choice and short, open-ended items. The expected information of each type of item was computed. On average, four-category polytomous items yielded 2.1 to 3.1 times as much IRT information as dichotomous items. These results provide limited support for the ad hoc rule of weighting "k" category polytomous items the same as "K-1" dichotomous items for computing total scores. Comparing average values, polytomous items provided more information across the entire proficiency range and more information about examinees of moderately high proficiency. When scored dichotomously, information in the extended open-ended items sharply decreased. However, they still provided more expected information than did the other response formats. For reference, a derivation of the information function for the generalized partial credit model is included in an appendix. Four tables and five figures illustrate the analysis. (Contains 17 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-93-12

ED 385 580

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

AN EMPIRICAL EXAMINATION OF THE IRT INFORMATION IN POLYTOMOUSLY SCORED READING ITEMS

John R. Donoghue

Educational Testing Service
Princeton, New Jersey
March 1993

TM024018

BEST COPY AVAILABLE

An Empirical Examination of the IRT Information in Polytomously Scored Reading Items

John R. Donoghue

Educational Testing Service

The work on which this paper is based was performed for the National Center for Educational Statistics, Office of Educational Research and Improvement, by Educational Testing Service.

Copyright © 1993. Educational Testing Service. All rights reserved.

Abstract

One natural question about polytomous items (which yield responses which can be scored as ordered categories) concerns the information contained in the items; how much more information do polytomous items yield? Using the generalized partial credit IRT model, polytomous items from the 1991 field test of the NAEP Reading Assessment were calibrated with multiple choice and short open-ended items. The expected information of each type of item was computed.

On average, four-category polytomous items yielded 2.1-3.1 times as much IRT information as dichotomous items. These results provide limited support for the *ad hoc* rule of weighting *k* category polytomous items the same as *k-1* dichotomous items for computing total scores. Comparing average values, polytomous items provided more information across the entire proficiency range. Polytomous items provided the most information about examinees of moderately high proficiency; the information function peaked at 1.0 to 1.5, and the population distribution mean was 0. When scored dichotomously, information in the extended open-ended items sharply decreased. However, they still provided more expected information than did the other response formats.

For reference, a derivation of the information function for the generalized partial credit model is included.

An Empirical Examination of the IRT Information in Polytomously Scored Reading Items

Recent years have seen a growing use items which yield responses which can be scored as ordered categories. Such polytomous items typically require more testing time, and are more expensive to score than are dichotomous items. As a result, questions arise as to the effectiveness of polytomous items relative to that of dichotomous items. One natural question about polytomous items concerns the information contained in the items; how much more information do polytomous items yield? To date, there is little empirical data dealing with this issue. Wainer and Thissen (1992) used the classical test theory concept of reliability to examine the relative effectiveness of sections composed of polytomous and dichotomous items in the College Board's Advanced Placement Chemistry Exam. Comparing the reliability of the sections, they used the Spearman-Brown prophesy formula to determine that many polytomous items would be required to yield the same reliability as the multiple choice section. Wainer and Thissen point out that, in terms of both time and expense, constructing such a test would be impractical. They conclude that polytomous items of the type they examined are inefficient, and question their utility.

Another approach to evaluating polytomous items is to use an item response theory (IRT) information based approach. IRT models for polytomous data have been around for quite some time. The graded response model goes back to the 1960s

(Samejima, 1969, 1972). Extensions of the Rasch model to polytomous items are almost as old. Andrich's rating scale model (1978) and Masters' partial credit model (1982) are prime examples. Recently, Muraki (1992), has extended the partial credit model, incorporating a separate slope parameter a_j for each polytomous item and using the EM algorithm to estimate model parameters by the method of marginal maximum likelihood. (Bock and Aitkin, 1981). The PARSCALE program (Muraki & Bock, 1991) allows the model parameters to be estimated.

Yamamoto and Kulick (1992) examined data from the National Assessment of Educational Progress (NAEP, see below) 1990 Science Trend Assessment. A small number of the items on the Science Assessment were constructed response items. Although these items were not intended to be used as polytomous items, ordered category scores were available for several of them. Using the NAEP version of the PARSCALE program, they scaled these items polytomously, and computed the relative information function for dichotomous and polytomous science items. They found that the polytomous items contained, on average, slightly less information than did the dichotomous items. They point out, however, that the items were not intended to be scored polytomously. Thus, it is not clear to what extent their findings are applicable to other polytomous items intended to be scored polytomously.

This paper provides an examination of the IRT-based information of polytomous items which were developed with the intention of polytomous scaling.

Method

Data

This study used data from the 1991 field test of the 1992 NAEP Reading Assessment. NAEP is a federally mandated survey of what American students at Grades 4, 8, and 12 know and can do. The NAEP contract is conducted by Educational Testing Service under the direction of the National Assessment Governing Board, and administered by National Center on Educational Statistics.

The 1992 Reading Assessment is a new assessment based on new objectives. The specifications were developed by a panel of reading experts using a consensus process. The assessment contains longer reading passages which are intended to be more authentic examples of the reading tasks encountered in and out of school. In addition to multiple choice items, each passage is followed by a number of constructed response items, accounting for approximately 40% of the testing time. Some of these items are relatively short open-ended items, requiring a few sentences or a paragraph response. These short open-ended items are typically scored as correct or incorrect. In addition, each reading passage contains at least one extended open-ended item, which requires a more in-depth, elaborated response. These extended open-ended items were scored polytomously:

0 - Unsatisfactory;

1 - Partial;

2 - Essential;

3 - Extensive, which demonstrate more in-depth understanding.

Detailed scoring rubrics were developed for each item. The actual items are secure, and so cannot be reproduced here. However, a typical extended open-ended item might ask the examinee to compare and contrast two accounts of a historical event, or to describe the feelings of a character in a story and describe the events in the story which triggered those feelings.

NAEP uses a balanced incomplete block (BIB) design. Separately timed sections, termed blocks, are combined to form booklets according to the BIB design. The individual booklets are spiralled, i.e., assigned to examinees according to a systematic arrangement such that each booklet is presented to a randomly equivalent group of examinees (see Messick, Beaton and Lord, 1983 for more details). To assess the proficiency of a population and important subgroups, BIB spiralling is very efficient; it allows a large number of items to be presented, while simultaneously limiting the testing time for an individual examinee. However, relatively little information is obtained for individual examinees. NAEP uses IRT to pull together the pieces of the BIB spiral assessment, to establish vertical (cross-grade) scales, and to perform trend analyses.

IRT Models

Student responses to the field test data were scaled using item response theory (IRT) methods. Multiple choice items were fit using a 3PL model:

$$P_j(\theta) = c_j + \frac{1-c_j}{1 + \exp(-Da_j(\theta - b_j))} \qquad (1)$$

where θ is the proficiency which underlies the responses to the test items (i.e., "reading ability"),

c_j is the guessing parameter, and corresponds to the probability that a subject of very low proficiency will get the item correct,

b_j is the difficulty parameter,

a_j is the discrimination (slope) parameter, and

D is the scaling factor 1.7.

Short open-ended items were fit using a 2PL model, which is identical to equation (1) although the c-parameter is constrained to equal 0. Figure 1 gives a typical 3PL curve.

Insert Figure 1 about here

Polytomous items were fit using the generalized partial credit model, where the responses are scored as the integers 0, 1, ..., m_j. A basic relationship in polytomous IRT models is the item category response function (ICRF). This function, denoted $P_{jk}(\theta)$, describes the probability that an examinee of given ability θ will obtain score k on item j. Figure 2 show the ICRFs for a typical four-category item. Assume that Figure 2 represented a constructed response mathematics item which required three steps to successfully complete, and that the scoring rubric gave partial credit. The curve labeled

"P0" would then give the probability that an examinee completes none of the steps. Similarly, P1, P2, and P3 would show the probability of completing, respectively, one step, two steps, and all three steps (complete solution). As ability increases, the probability of no steps (P0) decreases, and the probability of one or more steps correspondingly increases.

Insert Figure 2 about here

The generalized partial credit model states that the form of the ICRFs is:

$$P_{jk}(\theta) = \frac{\exp\left\{\sum_{c=0}^{k} Da_j(\theta - b_{jc})\right\}}{\sum_{c=0}^{m_j} \exp\left\{\sum_{v=0}^{c} Da_j(\theta - b_{jv})\right\}} \qquad (2)$$

where θ is the proficiency which underlies the responses to the test items (i.e., "reading ability"),

b_{jv} is the transition parameter, and denotes the ability for which scores k and $k-1$ are equally likely.

a_j is the discrimination (slope) parameter, and

D is the scaling factor 1.7.

By convention, d_{j0} is arbitrarily set to 0.0 (see Muraki, 1992).

The IRT information is a function of proficiency. The information for item j depends upon the model. For dichotomous models, the relations are well known (e.g.,

Hambleton and Swaminathan, 1985). The information in item j for the 3PL model is:

$$I_j(\theta) = \frac{D^2 a_j^2 (1-P_j(\theta))}{P_j(\theta)} \cdot \frac{(P_j(\theta)-c_j)^2}{(1-c_j)^2} \quad . \qquad (3)$$

Information in the 2PL model is obtained by setting c_j in Equation (3) to 0. For polytomous items in the generalized partial credit model, the information in item j is:

$$I_j(\theta) = D^2 a_j^2 \left[\sum_{k=0}^{m_j} k^2 P_{jk}(\theta) - \left(\sum_{k=0}^{m_j} k P_{jk}(\theta) \right)^2 \right] \quad . \qquad (4)$$

(Because this equation is not readily available, a derivation is included in the Appendix.) It is interesting to note that, for polytomous items, $I_j(\theta)$ can be viewed as a conditional variance. If the *k* values are treated as category scores, $I_j(\theta)$ is $D^2a_j^2$ times the variance of X_j, conditional on θ, sometimes written $\sigma^2(X_j|\theta)$.

Under the IRT assumption of local independence, the total information function for a group of n items is simply the sum of the item information functions:

$$I(\theta) = \sum_{j=1}^{n} I_j(\theta) \quad . \qquad (6)$$

Analysis

The data for the field test of the NAEP Reading Assessment were calibrated

according to the IRT model. Multiple choice items were fit using a 3PL model, short open-ended items were fit using a 2PL model, and extended open-ended items were fit using the generalized partial credit model. Data for each grade were scaled separately. A single, unidimensional scale was fit at each grade. Items which were not reached were treated as if they were not presented. Omitted responses were treated as fractionally correct for dichotomous items, and were combined with the lowest category (score of 0) for polytomous items.

The complete analysis of a data set consisted of the following steps. First, item responses were calibrated using PARSCALE (Muraki & Bock, 1991); the NAEP modified (Rogers, 1992) version of the program was used. Starting values were computed from item statistics based on the entire data set. PARSCALE calibrations were done in two stages. At stage one, the program was run to convergence by specifying a N(0,1) prior distribution of proficiency. The values of the item parameters from this normal solution were then used as starting values for a second stage estimation run in which the proficiency distribution (modelled as a multinomial distribution, on a fixed number of proficiency values, termed quadrature points) was estimated concurrently with item parameters (e.g., Mislevy & Bock, 1983). This two stage procedure was used for the 1990 NAEP assessments in Reading (Donoghue, 1992), Mathematics (Mazzeo, 1991; Yamamoto & Jenkins, 1992), and Science (Allen, 1992), and has proven effective in avoiding problems of local optima in the likelihood function.

The expected information was then computed for each item. This expectation was

based on the posterior distribution of proficiency, provided by PARSCALE. To do this, the information function for each item was evaluated at each of the quadrature points. The function was then multiplied by the posterior weight associated with that quadrature point, summed to yield the expected information for each item:

$$E_j(I) = \sum_{q=1}^{Q} w_q \cdot I_j(x_q) \quad . \qquad (7)$$

Table 1 summarizes the expected information for each type of item.

Insert Table 1 about here

Table 2 gives the relative expected information for polytomous items. This is the ratio of average expected information for polytomous items, divided by the average expected information for each type of dichotomous items. The ratio of the expected information of a polytomous to a multiple choice item ranged from 2.3 to 3.7, indicating that a typical polytomous item yields about two and one-third to three and two-thirds as much information as a typical multiple choice item. For short open-ended items, the ratio was somewhat smaller, ranging from 1.8 to 2.6. Compared to all dichotomous items, the extended open-ended items yielded 2.1 to 3.1 times as much information as dichotomous items.

Insert Table 2 about here

Figures 3 through 5 give the total information function for each type of item, normalized by the number of items. Thus, Figures 3 through 5 represent the average information function per item. The information functions for the polytomous items attain their maximum for higher proficiencies than do those for the short open-ended or multiple choice items. Polytomous items provided the most information about examinees of moderately high proficiency; the information function peaked at 1.0 to 1.5, compared to a population proficiency distribution mean of 0.

Insert Figures 3-5 about here

In attempt to further characterize these items, each of the polytomous items was dichotomized. The purpose of this analysis was to try to separate effects of having multiple score points (polytomous scoring) from those quality of the questions; are the polytomous items simply better items, or does the polytomous scoring add information? Each item was rescored to indicate whether or not the response provided the material essential to completely answer the question. Thus, responses scored 0 or 1 were treated as incorrect, and responses which received 2 or 3 were treated as correct. Each grade's data were again calibrated (using the procedures described above), with the dichotomized responses in place of the polytomous responses. Dichotomized items were calibrated using a 2PL model.

Table 3 gives descriptive statistics of the expected information for each item type. Because they are based on a different calibration, the numbers in Table 3 are not

directly comparable to those in Table 1. However, the relative information, given in Table 4, is comparable to the information in Table 2.

Insert Tables 3 and 4 about here

As would be expected, dichotomizing the extended open-ended items reduced the amount of information that they provided. The entries in Table 2 are .65 to 1.5 higher than the corresponding entries in Table 4. However, with one exception, the entries in Table 4 are greater than 1.0, indicating that, even when dichotomized, the extended open ended items still provide more information than either multiple choice or short open-ended items. This is more true at the higher grades.

Conclusions

Using data from the 1991 field test of the NAEP Reading Assessment, polytomous items were found to yield substantially more information than did dichotomous items; ratios of expected information range from 2.1 to 3.1. These results indicate that, for these data at least, polytomously scored constructed response items may provide a substantial increase in the information per item.

The results obtained are in some ways contrary to those of Yamamoto and Kulick (1992) and Wainer and Thissen (1992). The differences with Yamamoto and Kulick may be due to several factors. The items and scoring rubrics used in this study were

developed to be scored polytomously. The data examined by Yamamoto and Kulick were developed knowing that they would be scored dichotomously. The ordered categories were used to further characterize student responses. It seems reasonable that the intentional nature of the test development process is an important part of constructing good polytomous items.

Part of the difference with the Wainer and Thissen (1992) study is the result of a difference in focus. Wainer and Thissen focused on testing time and expense; polytomously scored items do typically take longer and cost more than dichotomous items. The focus of this study is on information; polytomously scored items can yield substantially more information than an equal number of dichotomous items. This must be considered as one more factor in the debate over such items. Thus, it is a factor to combine with concerns of validity and "authenticity," and balanced against concerns of cost effectiveness and ease of development and scoring.

Finally, the results obtained here do provide some support for the common, *ad hoc*, procedure of scoring polytomous items from 0 to *k-1*, effectively weighting them as *k-1* dichotomous items. For this data set, this value was slightly too small when compared to multiple choice items, and somewhat too large when compared to short-opened items. These results indicate that the procedure provides a reasonable approximation. Thus, the procedure may yield quite satisfactory results when forming total scores, such as for item analyses.

References

Allen, N. L. (1992). "Data Analysis for the Science Assessment." In E. G. Johnson & N. L. Allen (Eds.), The NAEP 1990 technical report, Report 21-TR-20. Washington, D. C.: National Center for Educational Statistics.

Andrich, D. (1978). A rating formulation for ordered response categories. Psychometrika, 43, 561-573.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46, 443-459.

Donoghue, J. R. (1992). "Data Analysis for the Reading Assessment." In E. G. Johnson & N. L. Allen (Eds.), The NAEP 1990 technical report, Report 21-TR-20. Washington, D. C.: National Center for Educational Statistics.

Hambleton, R. K. & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer-Nijhoff Publishing.

Masters, G. M. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149-174.

Mazzeo, J. (1991). Data analysis and scaling. In S. L. Koffler (Ed.), The technical report of NAEP's 1990 Trial State Assessment Program 20-ST-01. Washington, D. C.: National Center for Educational Statistics.

Messick, S., Beaton, A. & Lord, F. (1983). National Assessment of Educational Progress reconsidered: A new design for a new era. NAEP Report 83-1. Princeton, NJ: Educational Testing Service.

Mislevy, R. J. & Bock, R. D. (1983). BILOG: Item analysis and test scoring with binary logistic models [Computer program]. Mooresville, IN: Scientific Software, Inc.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16, 159-176.

Muraki, E. & Bock, R. D. (1991). PARSCALE: Parameter scaling of rating data [computer program]. Chicago, IL: Scientific Software, Inc.

Rogers, A. (1992). PARSCALE: Parameter scaling of rating data [Computer program, modified]. Princeton, NJ: Educational Testing Service.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement, 34 (Monograph Number 17).

Samejima, F. (1972). A general model for free-response data. Psychometrika Monograph Supplement, 37 (Monograph Number 18).

Wainer, H. & Thissen, D. (1992). Combining multiple-choice and constructed response test scores: Toward a Marxist theory of test construction. Program Statistics Research Technical Report RR-92-23. Princeton, NJ: Educational Testing Service.

Yamamoto, K. & Jenkins, F. (1992). "Data Analysis for the Mathematics Assessment." In E. G. Johnson & N. L. Allen (Eds.), The NAEP 1990 technical report, Report 21-TR-20. Washington, D. C.: National Center for Educational Statistics.

Yamamoto, K. & Kulick, E. (1992). An information-based approach to maintaining content validity and determining the relative value of polytomous and dichotomous items. Paper presented at the annual meeting of the American Educational Research Association, April, 1992, San Francisco, CA.

Table 1

Descriptive Statistics for
Expected Information of Different Item Types

	N	Mean	Min.	25th %tile	Median	75th %tile	Max.
Age 9							
Multiple Choice	114	0.236	0.018	0.139	0.208	0.322	0.788
Short Open-Ended	70	0.310	0.053	0.192	0.300	0.403	0.839
Extended Open-Ended (Polytomous)	16	0.552	0.247	0.404	0.501	0.691	1.165
Age 13							
Multiple Choice	108	0.157	0.008	0.099	0.134	0.209	0.481
Short Open-Ended	73	0.221	0.029	0.136	0.210	0.309	0.445
Extended Open-Ended (Polytomous)	20	0.576	0.201	0.282	0.480	0.756	1.535
Age 17							
Multiple Choice	118	0.155	0.004	0.087	0.131	0.207	0.415
Short Open-Ended	66	0.248	0.051	0.154	0.222	0.340	0.634
Extended Open-Ended (Polytomous)	21	0.491	0.138	0.282	0.483	0.613	0.964

Table 2

Relative Information for Different Item Types

	Grade 4	Grade 8	Grade 12
Item Type[1]	Ratio of Average Information	Ratio of Average Information	Ratio of Average Information
EOE/MC	2.34	3.67	3.17
EOE/SOE	1.78	2.61	1.98
EOE/D	2.09	3.15	2.61

[1] Abbreviations: MC - Multiple choice, SOE - Short open-ended, EOE - Extended open-ended, D - Dichotomous, i.e., multiple choice and short open-ended.

Table 3

Descriptive Statistics for
Expected Information for Different Item Types
Calibrated with Extended Open-ended Items Dichotomized

	N	Mean	Min.	25th %tile	Median	75th %tile	Max.
Age 9							
Multiple Choice	114	0.231	0.015	0.140	0.202	0.306	0.829
Short Open-Ended	70	0.302	0.055	0.197	0.296	0.391	0.744
Extended Open-Ended (Dichotomized)	16	0.295	0.108	0.180	0.288	0.386	0.611
Age 13							
Multiple Choice	108	0.164	0.006	0.102	0.141	0.216	0.530
Short Open-Ended	73	0.226	0.031	0.141	0.215	0.311	0.486
Extended Open-Ended (Dichotomized)	20	0.332	0.043	0.181	0.318	0.494	0.578
Age 17							
Multiple Choice	118	0.158	0.004	0.087	0.137	0.211	0.431
Short Open-Ended	66	0.253	0.051	0.159	0.220	0.355	0.586
Extended Open-Ended (Dichotomized)	21	0.337	0.099	0.178	0.329	0.477	0.710

Table 4

Relative Information for Different Item Types
Calibrated with Extended Open-ended Items Dichotomized

	Grade 4	Grade 8	Grade 12
Item Type[1]	Ratio of Average Information	Ratio of Average Information	Ratio of Average Information
DOE/MC	1.28	2.02	2.13
DOE/SOE	0.98	1.47	1.33
DOE/D	1.14	1.76	1.76

[1] Abbreviations: MC - Multiple choice, SOE - Short open-ended, DOE - Dichotomization of Extended open-ended, D - Dichotomous, i.e., multiple choice and short open-ended.

Figure 1

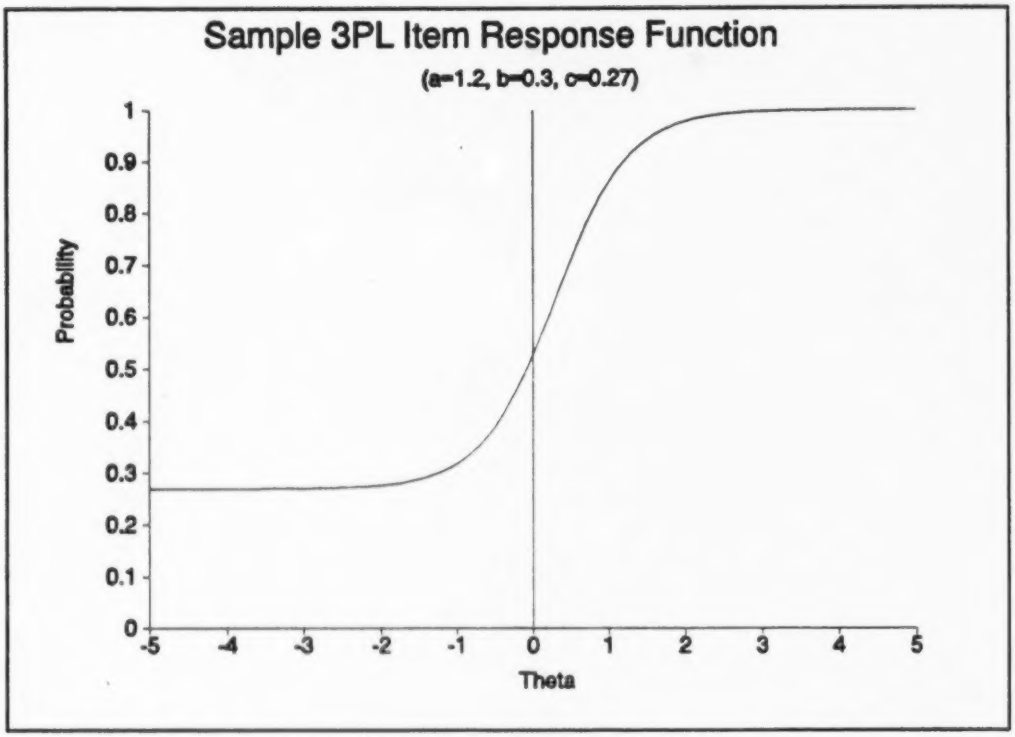

Sample 3PL Item Response Function
(a=1.2, b=0.3, c=0.27)
Probability
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
Theta

Figure 2

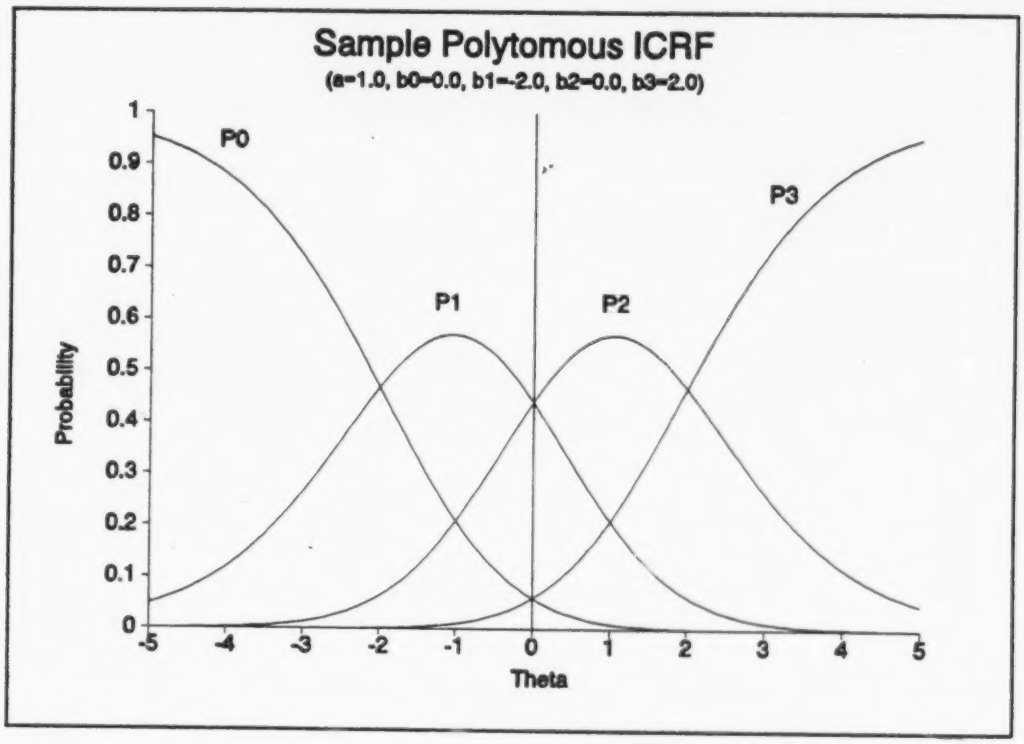

Sample Polytomous ICRF
(a=1.0, b0=0.0, b1=-2.0, b2=0.0, b3=2.0)
P0
P1
P2
P3
Probability
Theta
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5

Figure 3

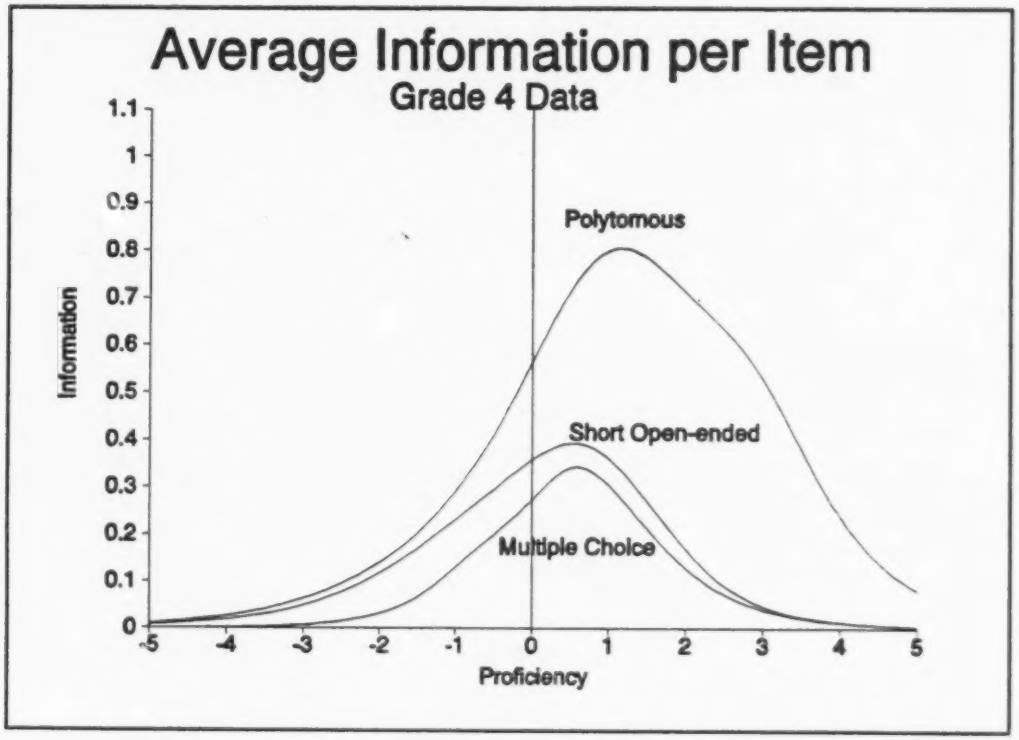

Average Information per Item
Grade 4 Data
Information
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
Proficiency
Polytomous
Short Open-ended
Multiple Choice

Figure 4

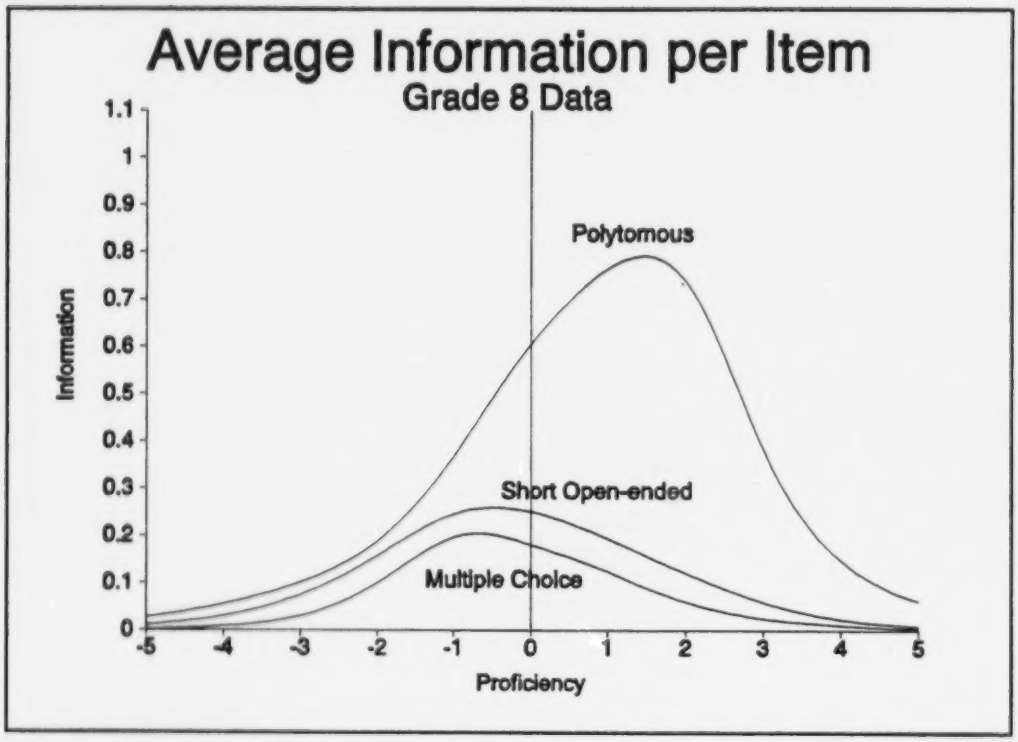
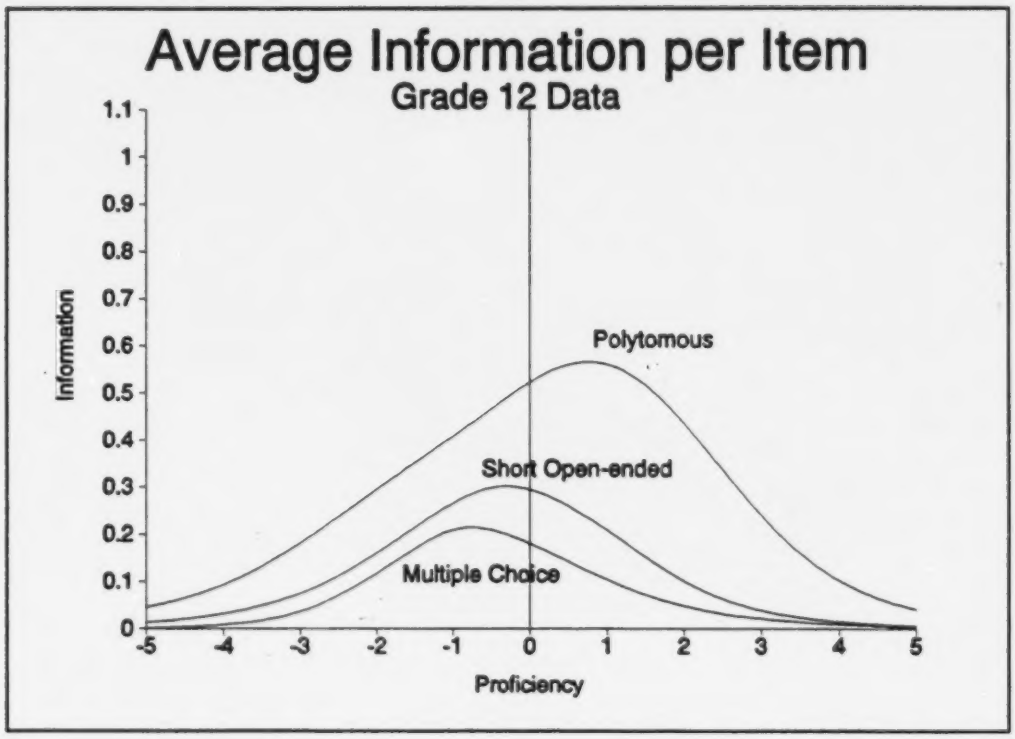

Average Information per Item
Grade 8 Data
Information
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Polytomous
Short Open-ended
Multiple Choice
-5
-4
-3
-2
-1
0
1
2
3
4
5
Proficiency

Figure 5


Average Information per Item
Grade 12 Data
Information
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Polytomous
Short Open-ended
Multiple Choice
-5
-4
-3
-2
-1
0
1
2
3
4
5
Proficiency

Appendix

Derivation of the Information Function for the Generalized Partial Credit Model

The information function $I(\theta)$ is defined:

$$I(\theta) = -E\left(\frac{\partial^2 L'(\theta)}{\partial \theta^2}\right) \tag{1}$$

where $L' = \ln L(\theta)$, i.e., the log likelihood evaluated at θ. The expectation may be seen as taken over subjects with fixed proficiency ($\theta = \theta_0$). To simplify notation, $P_{jk} = P_{jk}(\theta)$, $L = L(\theta;\underset{\sim}{P})$, $L' = \ln L(\theta;\underset{\sim}{P})$. For item j, with categories 0, 1, ..., m, the likelihood L is multinomial with parameter $\underset{\sim}{P} = (P_{j0}, P_{j1} \ldots P_{jm})$. If we define U_{jk} as an indicator function

$$U_{jk} = \begin{cases} 1, & x_j = k \\ 0, & otherwise \end{cases} \tag{2}$$

The likelihood is

$$L = \prod_{k=0}^{m} P_{jk}^{U_{jk}} \tag{3}$$

and L', the log likelihood is

$$L' = \sum_{k=0}^{m} U_{jk} \ln P_{jk} . \tag{4}$$

$$\frac{\partial L'}{\partial \theta} = \sum_{k=0}^{m} \frac{U_{jk}}{P_{jk}} \frac{\partial P_{jk}}{\partial \theta} \tag{5}$$

and

$$\frac{\partial^2 L'}{\partial \theta^2} = \sum_{k=0}^{m} \frac{-U_{jk}}{P_{jk}^2} \left(\frac{\partial P_{jk}}{\partial \theta} \right)^2 + \sum_{k=0}^{m} \frac{U_{jk}}{P_{jk}} \frac{\partial^2 P_{jk}}{\partial \theta^2} . \tag{6}$$

Under the generalized partial credit model,

$$P_{jk}(\theta) = \frac{\exp\left\{ \sum_{c=0}^{k} Da_j (\theta - b_{jc}) \right\}}{\sum_{c=0}^{m} \exp \sum_{v=0}^{c} Da_j (\theta - b_{jv})} . \tag{7}$$

By the quotient rule for derivatives

$$\frac{\partial P_{jk}}{\partial \theta} = \left\{ \left[(k+1)\, D\, a_j \exp \sum_{c=0}^{k} Da_j(\theta - b_{jc}) \right] \left[\sum_{c=0}^{m} \exp \sum_{v=0}^{c} Da_j(\theta - b_{jv}) \right] \right.$$

$$\left. - \left[\exp \sum_{c=0}^{k} D\, a_j (\theta - b_{jc}) \right] \left[\sum_{c=0}^{m} D\, a_j (c+1) \exp \sum_{v=0}^{c} D\, a_j (\theta - b_{jv}) \right] \right\}$$

$$/ \left[\sum_{c=0}^{m} \exp \sum_{v=0}^{c} Da_j (\theta - b_{jv}) \right]^2 \tag{8}$$

$$= Da_j P_{jk}\left[k - \sum_{c=0}^{m} cP_{jc}\right]. \tag{9}$$

$$\frac{\partial^2 P_{jk}}{\partial\theta^2} = Da_j \frac{\partial P_{jk}}{\partial\theta}\left[k - \sum_{c=0}^{m} cP_{jc}\right] + Da_j P_{jk}\left[-\sum_{c=0}^{m} c\frac{\partial P_{jc}}{\partial\theta}\right] \tag{10}$$

$$= D^2 a_j^2 P_{jk}\left[k - \sum_{c=0}^{m} cP_{jc}\right]^2 - D^2 a_j^2 P_{jk}\left[\sum_{c=0}^{m} cP_{jc}\left[c - \sum_{g=0}^{m} gP_{jg}\right]\right]. \tag{11}$$

Let $\lambda = \sum_{c=0}^{m} cP_{jc}$, and $\nu = \sum_{c=0}^{m} c^2 P_{jc}$. Note that $\lambda = E(c|\theta)$ and $\nu = E(c^2|\theta)$.

$$\frac{\partial^2 P_{jk}}{\partial\theta^2} = D^2 a_j^2 P_{jk}[(k-\lambda)^2 - \nu + \lambda^2] \tag{12}$$

$$= D^2 a_j^2 P_{jk}[k^2 - 2k\lambda + 2\lambda^2 - \nu]. \tag{13}$$

Substituting (9) and (13) into (6), we get

$$\frac{\partial^2 L'}{\partial\theta^2} = \sum_{k=0}^{m} \frac{-U_{jk}}{P_{jk}^2} D^2 a_j^2 P_{jk}^2 [k^2 - 2k\lambda + \lambda^2]$$

$$+ \sum_{k=0}^{m} \frac{U_{jk}}{P_{jk}} D^2 a_j^2 P_{jk} [k^2 - 2k\lambda + 2\lambda^2 - \nu]. \tag{14}$$

Canceling the P_{jk} and collecting like terms gives

$$= D^2a_j^2 \left[\sum_{k=0}^{m} U_{jk} (-k^2 + 2k\lambda - \lambda^2 + k^2 - 2k\lambda + 2\lambda^2 - v) \right] \quad (15)$$

$$= D^2a_j^2 \sum_{k=0}^{m} U_{jk} (\lambda^2 - v) . \quad (16)$$

Noting $E(U_{jk}|\theta) \equiv P_{jk}$

$$I_j(\theta) = -E\left(\frac{\partial L'}{\partial \theta^2}\right) = -D^2a_j^2 (\lambda^2 - v) \sum_{k=0}^{m} P_{jk} . \quad (17)$$

Because $\sum_{k=0}^{m} P_{jk} = 1$,

$$= D^2a_j^2(v - \lambda^2) , \quad (18)$$

$$I_j(\theta) = D^2a_j^2 \left[\sum_{k=0}^{m} k^2 P_{jk}(\theta) - \left(\sum_{k=0}^{m} kP_{jk}(\theta) \right)^2 \right] . \quad (19)$$

It is interesting to note that $I(\theta)$ can be viewed as a conditional variance. If the k values are treated as category scores, $I_j(\theta)$ is $D^2a_j^2$ times the variance of X_j, conditional on θ.

For a dichotomous item, we get

$$I_j(\theta) = D^2a_j^2[P_{j1} + 4P_{j2} - (P_{j1} + 2P_{j2})^2] .$$

$P_{j1} = 1 - P_{j2}$, where $P_{j2} = P_j$, the usual ICC for a dichotomous item

$$= D^2a_j^2[1 - P_j + 4P_j - (1 - P_j + 2P_j)^2]$$

$$= D^2a_j^2[1 + 3P_j - 1 - 2P_j - P_j^2]$$

$$= D^2a_j^2[P_j - P_j^2]$$

$$= D^2a_j^2P_j(1 - P_j)$$

which is the usual expression for the information of a 2PL item.